UNIVERSITY OF BELGRADE
FACULTY OF ORGANISATIONAL SCIENCES

40 let
Univerza v Mariboru

TEMPLE
UNIVERSITY

Towards predictive personalized preventive medicine

Data and knowledge driven approach

Milan Vukicevic, University of Belgrade, Faculty of Organizational Sciences,
Center for Business Decision Making, Belgrade, Serbia

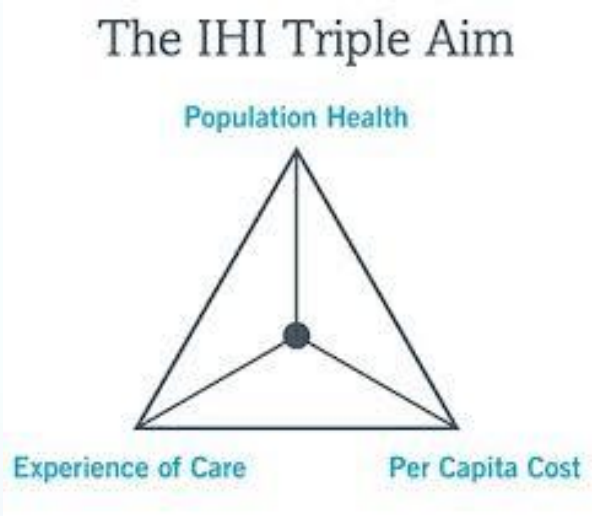
Acknowledgement

This research was supported by DARPA Grant FA9550-12-1-0406 negotiated by AFOSR, National Science Foundation through major research instrumentation, grant number CNS-09-58854, and by SNSF Joint Research project (SCOPES), ID: IZ73Z0_152415.

Predicted **high economic impact** of health care system on global economy, caused by severe chronic pathologies such as diabetes and cardiovascular disorders, led health care from reactive medicine to

(a) predictive medicine
(b) personalized, aiming to create treatment tailored to the person, and
(c) preventive.

- European Association for Predictive, Preventive and Personalized Medicine (EPMA),
- United Nations,
- European Union,
- National Institutes of Health.



Path: Primary care – Secondary care – Wellness promotion

How: Keeping patients out of hospitals, more engagement of the patients and their families in healthcare process

Where are we now

- Virtual visits (primary care doctors have easy access to specialists) = > Reduced re-admissions
- Connected Health (usage of connected devices for disease monitoring)
- Social Health platforms
  - Apple Healthcare
  - MicrosoftHealth, Healthvault.
  - The Oracle Health Management Platform Solution
  - Philips HealthSuite,
  - THE EMC Healthcare Analytics Solution
  - Patients like me etc.

Physicians are often overwhelmed and are not in a position to constantly monitor and process the large volumes of data that each patient generates

It is anticipated that in 2025 there will be a shortage of 90000 doctors in the US alone) and this is often an important reason for wrong or late diagnoses and care.

Challenges

Data fusion and analyses of heterogenous and multy-scale, dynamic sources of data (both spacial and temporal)
Data privacy
High complexity of the problems - complex diseases, multitude of factors that influence on development and progress of diseases like predispositions, lifestyle, disease history etc.,
Multi-modality of the data - e.g. Electronic Health Records - EHRs, data streams from wearables, medical image data, -Omix data, Questionaries etc.,
Partially observed data - analyises of only subset of all potentially available data (sympthoms, laboratory tests, EEG, X-rays etc.) which, is often the case since collection of all data can be expensive, but also contra-effective for patients heath can give partial view of patients health state and thus lead to uncertain/wrong decisions),
Context changes (e.g. change of some vital signs could be caused by increased physical activity or weather changes and not by some disease) etc.

Unreliable predictive algorithms, wrong decisions have high human and financial costs.
Predictive models try to learn such complex systems from scratch based only on data, while disregarding available domain knowledge and vice versa.

This leads to a situation where doctors and patients cannot benefit from data driven models and data driven models do not exploit existing domain knowledge, leaving a **large gap between actual data usage and potential data usage** in healthcare that prevents a paradigm shift from delayed interventional to predictive person-tailored medicine.

PPPM – SHRECK

Predictive, Preventive, and Personalized Medical Platform based on Smart Health Records and Enriched Clinical Knowledge

Knowledge sources

Ontologies

Expert systems

Rules

Data

Blood pressure

Activity

EHRs

Heart rate

Sugar level

EEG

Weight

Nutrition

Knowledge & Data Fusion

Knowledge model

Preventive recommendations

Therapy recommendations

Disease risk

Static Context:

Demographics

Disease history

Predispositions

Job

Dynamic Context:

Weather conditions

Symptoms

Epidemics

Extreme efforts

Advanced context aware algorithms

Structured learning

Multi – task learning

Near real time stream analyses

Early classification and regression

Anomaly detection and prediction

Deliverables

Collaborative PPPM Platform

Environment for knowledge sharing and evaluation of predictive models

APIs for development and integration and of advanced predictive algorithms into healthcare applications

Applications

Therapy optimization

Diagnostics

Health state screening

Therapy recommendation

Early warning systems

Risk assessment

Stakeholders:

Patients

Clinicians

Insurance companies

Hospitals

Wearable device companies

Healthcare application developers

Impact

Improved decision support in healthcare

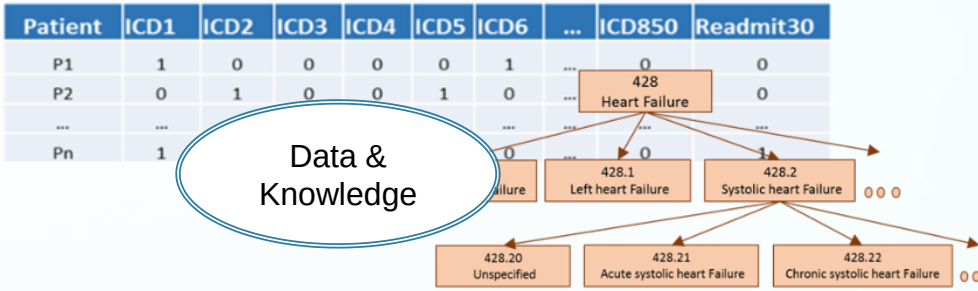
Stimulating patient participation in health care

Stimulating development of PPPM apps

GHFCS: ICD-9 based Feature Space Compression for 30-day Hospital Re-admission prediction

Predicting Hospital Re-admissions - high impact on improvement of healthcare services and reducing costs

- **Challenge:** high sparsity and dimensionality of data.
- **Objective:** develop efficient feature compression
  - Interpretable predictive models
  - no loss in predictive performance.



The idea: network compression – aggregate data based on ICD-9 hierarchical graph

GHFCS (Group hierarchical feature compression and selection)

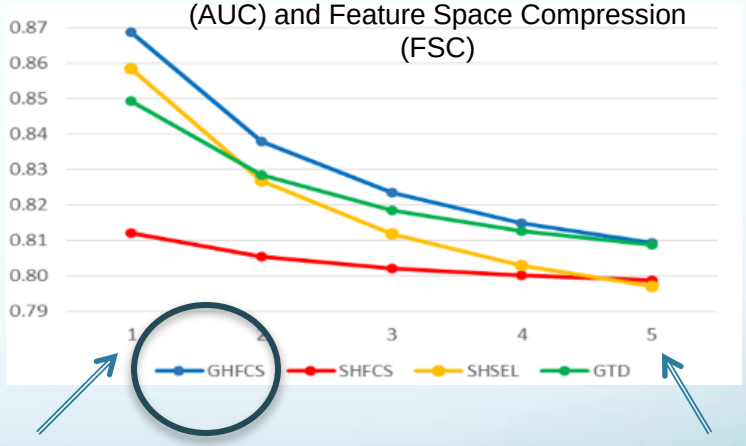
- aggregates features on the highest levels possible (without loss in information potential)
- Allows comparison of feature information potential on all ICD-9 hierarchical levels and paths.

Application: re-admission prediction for pediatric patient data (HCUP) from CA
851 features on the lowest level of hierarchy

Results :

- Traditional methods reduce feature space, but result in significant loss in predictive accuracy
- **GHFSC** gave the most interpretable solution (20 features) without loss predictive performance compared to similar methods
- Multi-scale learning

Harmonic mean between Area Under Curve (AUC) and Feature Space Compression (FSC)



AUC and FSC are equally important

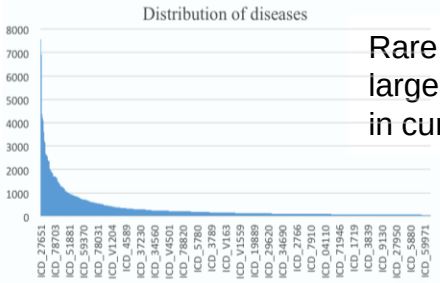
AUC is 5 times more important than FSC

Radovanovic, S, Vukicevic, M, Kovacevic, A, Sliglic, G, Obradovic, Z. (2015) "Domain knowledge based hierarchical feature selection for 30-day hospital readmission prediction", Proc. AIME 2015, the 15th Conference on Artificial Intelligence in Medicine, Pavia, Italy, June, 2015.

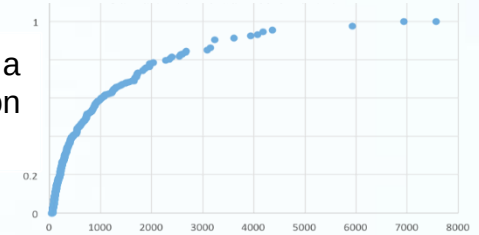
ICD9-VEG: EHR and ICD-9 based learning for Re-admission Risk Prediction

Problem: 30-day Hospital re-admission prediction
Data: pediatric patients from CA (HCUP)

Most of diseases are rarely observed.

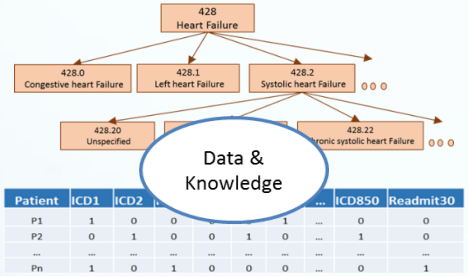


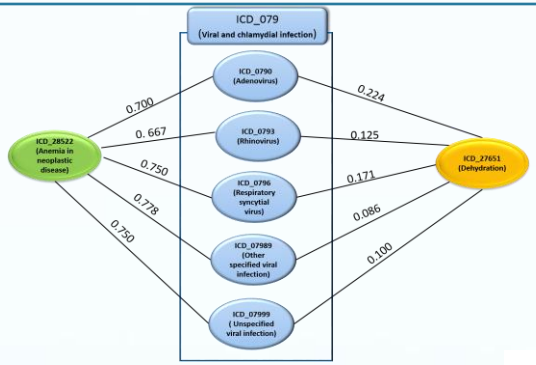
Rare diseases constitute a large portion of re-admission in cumulative.



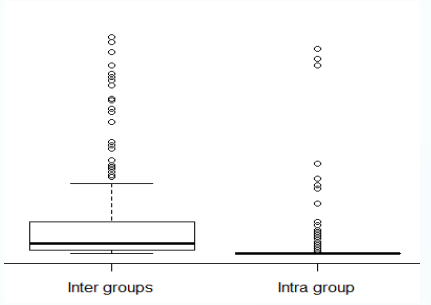
X-axis - diseases in ascending order by frequency of appearance
Y-axis - cumulative share of each disease in total number of readmission)

The idea: use prior knowledge from ICD-9 ontology for randomization.





Some diseases have similar re-admission risks with whole ICD groups



Inter-groups comorbidities are more common then intra-group comorbidities

ICD-9-VEG tool for generating data and knowledge based Virtual Examples - uses randomization which is controlled by ICD-9 graph.
Application: Using ICD-9 VEG to generate virtual examples for rare diseases and comorbidities, thus removing bias of algorithm towards frequently observed ones.

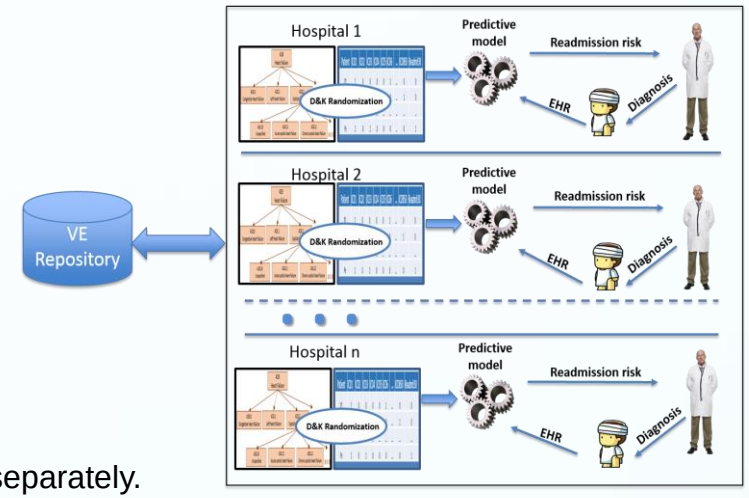
Experiment: Logistic regression was applied on data enriched with Virtual examples constructed by ICD-9 VEG. Performance is compared with several oversampling and ensemble techniques
Result: this strategy improves predictive performance and allows generation of unobserved comorbidities

Vukicevic, M, Radovanovic, S, Kovacevic, A, Stiglic G, Obradovic, Z. (2015) “Improving hospital readmission prediction using domain knowledge based virtual examples ”,Proc. KMO 2015, the 10th Conference on Knowledge Management in Organization, Maribor, Slovenia, August, 2015

Privacy Preserving DSS for reducing Hospital Re-admission rates based on predictive models and knowledge and data sharing

Lack of data is often the major obstacle for evolving highly accurate predictive models.
Reasons: rare diseases, long and expensive procedures for data collection and confidentiality of personally sensitive information.

Privacy preserving DSS for EHR of information about EHRs between hospitals, while preserving privacy (through common VE repository).
Prevention of data quality loss by randomization VE generator can use some domain knowledge source (ontologies or rules) in order to randomize the original data in controlled manner.



Experiments:
Original - LR was evaluated on the original data from each hospital separately.
Shared – LR model is created on data from all hospitals (simulation of situation where data could be shared).
Virtual Examples –Original data from each hospital is enriched with the data from VE repository.

Hospital	#patients	#readmitted	%readmission	Original	Shared	VE repository
1	7884	1,336	16.95	0.695	0.820	0.815
2	6394	1,450	22.68	0.693	0.793	0.771
3	6317	1,064	16.84	0.644	0.782	0.762
4	5103	705	13.82	0.621	0.780	0.794
5	4405	813	18.46	0.636	0.728	0.761

Conclusion: Knowledge based VE allow data sharing without loss in prediction accuracy.

Vukicevic, M, Radovanovic, S, Kovacevic, A, Delibasic B, Suknovic M (2015) “Privacy Preserving DSS forreducing Hospital Re-admission rates based on predictive models and knowledge and data sharing ”,Proc. of International Conference Decision Support Systems Technology , Belgrade, Serbia, May, 2015

Data

National Inpatient Sample (NIS) data - is the largest all-payer, uniform and multi-state inpatient care database that is publicly available in the United States. The archive is designed to approximate a 20-percent stratified sample of U.S. community hospitals. The utilized portion of the database, years 2003 to 2009over 56 million hospitalizations.

State inpatient Database SID data all-payer, uniform and state-specific inpatient discharge records, Demographic information (like age, birth year, sex, race), diagnosis (primary and up to 25 others), procedures (up to 25), information about hospitals and other information (like length of stay, total charges, type of payment and payer, discharge month, survival information)..

California state SID database contains 35,844,800 inpatient discharge records over 9 years (admissions from January 2003 to December 2011) for 19,319,350 distinct patients in 474 hospitals The SID database for New York state contains around 2.5 million records.

The Clinical Practice Research Datalink (CPRD) is consistent with including prescribed primary care drugs, administered hospital drugs, laboratory data, consultations, hospital coded disease data, disease registers and cancer registers,laboratory tests, pathology results, and lifestyle factors (height, weight, BMI, smoking andalcohol consumption, socioeconomic status, etc).

Health Retirement Study (HRS) data, 12545 samples (human individuals). There are data about 2.5 milion SNPs (single nucleotide polimorphisms) for that subjects (as measured, and up to 20 M of inputed ones...). Such genetic data are also accompanied with various phenotype information. These data are collected in longitudinal study that lasts for about 30 years, so it also have a **temporal** character.

Recent work

- Ghalwash, M., Radosavljevic, V, and Obradovic, Z. (2014) "Utilizing Temporal Patterns for Estimating Uncertainty in Interpretable Early Decision Making," *Proc. 20th ACM SIGKDD Conf. on Knowledge Discovery and Data Mining*, New York, NY, Aug. 2014. (PDF)
- Ghalwash, M., and Obradovic, Z. (2014) "A Data-Driven Model for Optimizing Therapy Duration for Septic Patients," *Proc. 14th SIAM Int'l Conf. Data Mining, 3rd Workshop on Data Mining for Medicine and Healthcare* , Philadelphia, April 2014. (PDF)a
- Mathew, G., Obradovic, Z. (2015) "A Distributed Decision Support Algorithm that Preserves Personal Privacy," *Journal of Intelligent Information Systems.*, Feb 2015, vol. 44, no. 1, pp. 107-132. (PDF)
- Ghalwash, M., Ramljak, D.,Obradovic, Z. (2015) "Patient-Specific Early Classification of Multivariate Observations," *International Journal of Data Mining and Bioinformatics*, Vol. 11, No. 4, 2015.
- Ramljak, D., Davey, A., Uversky, A., Roychoudhury, S., Obradovic, Z. (in press) "Casting a Wider Net: Data Driven Discovery of Proxies for Target Diagnoses," *AMIA 2015 Annual symposium,* San Francisco, Nov. 14 - 18 2015
- Vukicevic, M., Radovanovic, S., Kovacevic, A., Sliglic, G., Obradovic, Z. (2015) "Improving hospital readmission prediction using domain knowledge based virtual examples," *Proc. the 10th Conf. on Knowledge Management in Organization*, Maribor, Slovenia, August, 2015.
- Radovanovic, S., Vukicevic, M., Kovacevic, A., Sliglic, G., Obradovic, Z. (2015) "Domain knowledge based hierarchical feature selection for 30-day hospital readmission prediction" *Proc. AIME 2015, the 15th Conference on Artificial Intelligence in Medicine*, Pavia, Italy, June, 2015.
- Ramljak, D., Davey, A., Uversky, A., Roychoudhury, S., Obradovic, Z. (2015) "Hospital Corners and Wrapping Patients in Markov Blankets," *4th Workshop on Data Mining for Medicine and Healthcare, 2015 SIAM International Conference on Data Mining,* Vancouver, Canada, April 30 - May 02, 2015